CS6350 – Big Data Management and Analysis

Assignment submitted by Yash Vijaynarayan Gupta

Analysing Social Networks using GraphX/GraphFrame

Problem Statement - In this part, you will use Spark GraphX/GraphFrame to analyze social network data. You are free to choose any one of the Social network datasets available from the below link:
SNAP repository - https://snap.stanford.edu/data/index.html#socnets

You will use this dataset to construct a GraphX/GraphFrame graph and run some queries and algorithms on the graph. I have chosen the reddit database https://snap.stanford.edu/data/soc-RedditHyperlinks.html

Solution:

Load the data into a GraphFrame or RDD using Spark. Define a parser so that you can identify and extract relevant fields. Note that edges are directed, so if your dataset has undirected relationships, you might need to convert those into 2 directed relationships. That is, if your dataset contains an undirected friendship relationship between X and Y, then you might need to create 2 edges one from X to Y and the other from Y to X. Define edge and vertex structure and create property graphs.

Queries implemented with results:

Run the following queries using the GraphX/GraphFrame API and write your output to a file specified by the output parameter.

a. Find the top 5 nodes with the highest outdegree and find the count of the number of outgoing edges in each

	id	outDegree
1	subredditdrama	4665
2	circlebroke	2358
3	shitliberalssay	1968
4	outoftheloop	1958
5	copypasta	1824

b. Find the top 5 nodes with the highest indegree and find the count of the number of incoming edges in each

	id	inDegree
1	askreddit	7329
2	iama	3694
3	pics	2779
4	writingprompts	2490
5	videos	2446

c. Calculate PageRank for each of the nodes and output the top 5 nodes with the highest PageRank values. You are free to define any suitable parameters.

	id	pagerank
1	askreddit	592.0040787061722
2	iama	484.09984860204776
3	videos	312.0979902738428
4	pics	242.5159664733212
5	leagueoflegends	189.48743810454246

d. Run the connected components algorithm on it and find the top 5 components with the largest number of nodes.

	id	component
1	stephaniemichelle	1692217114753
2	ultimatepatreon	1692217114753
3	challenger	1632087572613
4	srt	1632087572613
5	lifepluslair	1614907703381

e. Run the triangle counts algorithm on each of the vertices and output the top 5 vertices with the largest triangle count. In case of ties, you can randomly select the top 5 vertices.

	count	id
1	31967	askreddit
2	26072	subredditdrama
3	24581	iama
4	15898	outoftheloop
5	11938	videos

Summary:

1. There are reddits and subreddits with two types of communities namely, attackers and defenders.
2. The top5 indegrees show popular sub reddits prone to negative (-1) and positive (1) reddits.
3. Similarly top 5-outdegrees show the top critics or top active users that comment on positive (1), negative (-1) subreddits.
4. PageRank gives top 5 popular nodes which in our case are reddits and similarly the top communities can be detected using connected Components () with most number of interrelated reddits.
5. The inter-relations between nodes is represented by triangle count just like in our case. Thus, the dataset makes sense when using graphframes.